Comp. Comp. & Maths. with Appls. Vol. 8, No. 3, pp. 217–229, 1982
Printed in Great Britain.

0097–4943/82/030217–13$03.00/0

INSTRUCTIVE EXPERIMENTS WITH SOME RUNGE–KUTTA–ROSENBROCK METHODS

JAN G. VERWER
Mathematical Centre, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

(*Received June* 1981; *in revised form August* 1981)
Communicated by L. F. Shampine

Abstract—The paper deals with certain boundedness properties of Runge–Kutta–Rosenbrock methods when applied to nonlinear stiff systems. It reports some instructive examples and numerical experiments performed with a number of simple 2-stage schemes and the Rosenbrock code ROW4A. Attention is paid to the conversion of non-autonomous problems to the autonomous form. An important conclusion is that this conversion may lead to a significant loss in accuracy.

1. INTRODUCTION

A substantial part of the literature on numerical methods for stiff systems of ordinary differential equations deals with Runge–Kutta–Rosenbrock methods. For the *non-autonomous* initial value problem

$$\dot{X} = F(t, X), \qquad X(t_0) = X_0, \tag{1.1}$$

the *original* m-stage Rosenbrock method (see [1]) is very similar to the Runge–Kutta integration formula

$$\begin{aligned} &X_n^{(0)} = X_n, \\ &K_n^{(j)} = [I - \gamma_j \tau J_n^{(j)}]^{-1} F(t_n^{(j)}, X_n^{(j)}), \qquad \gamma_j > 0,\ j = 0(1)m-1, \\ &X_n^{(j)} = X_n + \tau \sum_{l=0}^{j-1} \lambda_{j,l} K_n^{(l)}, \qquad j = 1(1)m, \\ &X_{n+1} = X_n^{(m)}, \qquad n = 0, 1, \ldots. \end{aligned} \tag{1.2}$$

X_n denotes the approximation at time $t = t_n$ and $\tau > 0$ denotes the stepsize; $t_n^{(j)} = t_n + \nu_j \tau$, where, normally, $0 \le \nu_j \le 1$. Further

$$\begin{aligned} &J_n^{(j)} = J(\hat{t}_n^{(j)}, \hat{X}_n^{(j)}), \qquad J(t, X) = \partial F(t, X)/\partial X, \\ &\hat{t}_n^{(j)} = \sum_{l=0}^{j} \alpha_{j,l} t_n^{(l)}, \qquad \hat{X}_n^{(j)} = \sum_{l=0}^{j} \alpha_{j,l} X_n^{(l)}, \end{aligned} \tag{1.3}$$

where the parameters $\alpha_{j,l}$ are real scalars. Note that each stage j involves an $F(t, X)$-evaluation, a solution of a system of linear algebraic equations, and, possibly, a $J(t, X)$-evaluation.

Up to now the literature on Rosenbrock schemes has mainly dealt with the development of new schemes and, in particular, with the analysis of their rational stability functions. In fact, it is now well-known that there do exist A-stable, or L-stable, Rosenbrock schemes of high order of consistency. It is less known, however, that such a scheme, which according to the Dahlquist–Henrici theory ought to be judged as being reliable, may behave very badly when applied to certain classes of non-linear problems. Or, when we are given two schemes of the same order of consistency and having the same stability function, we may encounter large differences in their performance when applied to these non-linear problems.

The present paper deals with these phenomena. We discuss a number of instructive examples and numerical experiments, most of which are based on results presented in a previous paper [2]. In that paper the author investigated, following ideas put forward by Stetter [3] and Van Veldhuizen [4], a so-called *uniform boundedness property* of method (1.2)

for two model classes which are directly relevant to non-linear stiff problems. This boundedness property plays a key role in the examples and experiments we are going to discuss.

In short, the contents of the paper are as follows. In Section 2 we briefly discuss the boundedness property we are concentrating on. Sections 3 and 4 review the model classes we investigated in [2]. In these sections we also discuss numerical examples. Section 5 deals with the conversion to the autonomous form which in the greater part of the literature is used whenever a genuine non-autonomous problem is met. An important conclusion of Section 5 is that this conversion to the autonomous form may lead to a significant loss in accuracy, and even to instability. In Section 5 we also report an experiment with the Rosenbrock code ROW4A[5]. Here our aim is to illustrate how bad boundedness properties show up in practice when using an automatic code.

2. THE PROPERTY OF ϵ-BOUNDEDNESS

In the analysis of numerical methods for stiff problems the study of *model-equations* has proved to be fruitful. For example, the simple well-known scalar model

$$\dot{x} = \delta x, \qquad \delta \in C, \qquad Re(\delta) < 0, \tag{2.1}$$

provides indispensable information on the absolute stability of integration methods for ordinary differential systems. For constant coefficient linear systems this scalar model already yields enough insight. For *non-linear stiff systems* however, this model appears to be too simple and there is a need for additional research on more refined models. Such a model (cf. [3, 4]) should permit the simultaneous *occurrence of smooth and transient solution components*, and its Jacobian matrix should have a *time-dependent eigensystem.* Further it should be possible to consider a limit process by which one can introduce *arbitrarily high stiffness.* Finally, the *occurrence of non-linear terms* in the model could increase our insight.

It is our purpose to support these views for Rosenbrock methods by means of some instructive examples and numerical experiments. Most of these will be based on theoretical results presented in [2]. There we investigated a so-called property of uniform boundedness for method (1.2) when applied to two model classes having the characteristics just mentioned. We shall now describe the kind of boundedness we have in mind. Let

$$\dot{X} = F(t, X, \epsilon), \qquad \epsilon \in (0, \epsilon_0], \qquad \epsilon_0 \text{ constant}, \tag{2.2}$$

represent some class of model equations, where (a) $t \in [t_0, T]$, t_0 and T finite and constant, $X(t_0) = X_0 = X_0(\epsilon)$. (b) All problems in this class possess a unique bounded solution $X = X(t, \epsilon)$ on $[t_0, T] \times (0, \epsilon_0]$, i.e. we suppose the existence of a constant K such that

$$\sup_{\epsilon \in (0, \epsilon_0]} \sup_{t \in [t_0, T]} \|X(t, \epsilon)\| \leq K.$$

(c) The stiffness ratio tends to infinity as $\epsilon \to 0$ ($1/\epsilon$ factors).

Note that the initial vector X_0 may depend on the stiffness parameter ϵ. This case may be relevant when we have non-linearities in X. In what follows it is convenient to represent the scheme (1.2) in the operator form

$$\begin{aligned} &X_n^{(0)} = X_n, \\ &X_n^{(j)} = \Phi^{(j)}(\{t_n^{(l)}, X_n^{(l)}\}_{l<j}, \tau, \epsilon; F), \qquad j = 1(1)m, \\ &X_{n+1} = X_n^{(m)}. \end{aligned} \tag{2.3}$$

DEFINITION 2.1

Suppose we are given a method of type (2.3) and a class of stiff problems satisfying properties (2.2a–c). We then call this method ϵ-bounded on this class if for all its problems the

following statement holds: for any point (t, X) in the region of definition of F, where $X = X(\epsilon)$ is bounded in $\epsilon \in (0, \epsilon_0]$, a constant τ^* exists such that

$$\Phi^{(j)}(\{t^{(l)}, X^{(l)}\}_{l<j}, \tau, \epsilon; F) = 0(1), \qquad \epsilon \to 0, \qquad j = 1(1)m, \tag{2.4}$$

for all $\tau \in (0, \tau^*]$, τ^* being independent of ϵ. □

For clarity we wish to make two comments on this definition. Firstly, in relation (2.4) we confine ourselves to fixed τ-values, i.e. the constant implied may depend on τ (cf. [3], p. 192). In view of property (2.2b), our goal is to select methods which, for a fixed value of τ, are able to produce a sequence of approximations over the interval $[t_0, T]$ which are bounded in $\epsilon \in (0, \epsilon_0]$. If, for a given problem, some sequence is unbounded as $\epsilon \to 0$, we may expect large discretization erros in a non-limit situation.

Our second comment concerns the additional boundedness requirement for $j < m$. We prefer to define ϵ-boundedness in this way as it facilitates the analysis (see [2]) and, of course, it is also natural to ask for boundedness of $\Phi^{(j)}$, $j < m$, if $\Phi^{(m)}$ is required to be bounded (in general $\Phi^{(m)}$ depends in a non-linear way on $\Phi^{(j)}$, $j < m$).

3. MODEL CLASS 1

In order to obtain concrete results on ϵ-boundedness one has to select appropriate model classes. In [2] we investigated two such classes. The first of these is reviewed in Section 3.1. In Section 3.2 we present a specific example to be used in Section 3.3 for a numerical illustration.

3.1 *A class of non-linear model equations*

The class is described by two coupled singularly perturbed differential systems of the form (see also [6])

$$\begin{aligned} \dot{x} &= f(t, x, y, \epsilon) + \epsilon^{-1} A(t) y, \qquad x(0) = x_0, \\ \dot{y} &= g(t, x, y, \epsilon) + \epsilon^{-1} \mu(t) B y, \qquad y(0) = y_0. \end{aligned} \tag{3.1}$$

We consider (3.1) on the interval $[0, T]$ and, until further notice, x_0, y_0 are assumed to be independent of ϵ. The functions on the r.h.s. are supposed to be sufficiently differentiable. The vector function f and g are allowed to be non-linear and they are supposed to be bounded in ϵ as $\epsilon \to 0$. Further, f: $[0, T] \times R^{s_1} \times R^{s_2} \times (0, \epsilon_0] \to R^{s_1}$ and g: $[0, T] \times R^{s_1} \times R^{s_2} \times (0, \epsilon_0] \to R^{s_2}$, where $s_1, s_2 \geq 1$. A is a t-dependent (s_1, s_2)-matrix and μ is a scalar function which is strictly positive, i.e. $\mu(t) \geq \tilde{\mu} > 0$ for all $t \in [0, T]$. Finally, B is a constant (s_2, s_2)-matrix whose spectrum $\Lambda(B)$ lies in the negative half plane $\mathbb{C}^- = \{z | Re(z) < 0\}$. It is not difficult to prove the following result [2]:

THEOREM 3.1

Let $\alpha = \max\{Re(\lambda)\colon \lambda \in \Lambda(B)\} < 0$. *Then, for all* $t \in (0, T]$ *and* $\epsilon \in (0, \epsilon_0]$, *the solution functions* $x(t, \epsilon)$ *and* $y(t, \epsilon)$ *of problem* (3.1) *satisfy*

$$\begin{aligned} &\|x(t, \epsilon)\| \leq K_0, \qquad \|\dot{x}(t, \epsilon)\| \leq K_1 \left[\epsilon^{-1} \exp\left(\tfrac{1}{2} \alpha \tilde{\mu} \epsilon^{-1} t\right) + 1\right] \\ &\|y(t, \epsilon)\| \leq \tilde{K}_0 \left[\exp\left(\tfrac{1}{2} \alpha \tilde{\mu} \epsilon^{-1} t\right) + \epsilon\right], \\ &\|\dot{y}(t, \epsilon)\| \leq \tilde{K}_1 \left[\epsilon^{-1} \exp\left(\tfrac{1}{2} \alpha \tilde{\mu} \epsilon^{-1} t\right) + 1\right], \end{aligned} \tag{3.2}$$

K_0, $\tilde{K}_0$, K_1 *and* $\tilde{K}_1$ *being positive constants independent of* t *and* ϵ. □

These inequalities reveal that we can write

$$x(t, \epsilon) = 0(1), \qquad y(t, \epsilon) = 0(\epsilon), \qquad \epsilon \to 0, \qquad t \in (0, T]. \tag{3.3}$$

Normally the x-solution will consist of a rapidly decaying transient component and a smooth one which determines $x(t, \epsilon)$ everywhere outside the transient phase. The transient behaviour of $x(t, \epsilon)$ is completely determined by the transient of the y-solution. Further, to a large extent the magnitude of the smooth component is independent of the stiffness parameter ϵ. For the y-solution the situation is somewhat different. Typically it contains a transient component and a smooth one which is $0(\epsilon)$ for all $t \in (0, T]$. In many practical situations it will be the smooth x-solution in which we are mostly interested, ϵ being so small that the transients can be neglected and that the smooth y-solution is of less practical interest. It will be clear now that an integration method suitable for (3.1) should generate approximations which show a similar behaviour in ϵ. In particular, the method should be capable of generating such approximations with some stepsize τ independent of ϵ, i.e. $X_n^{(j)} = [x_n^{(j)}, y_n^{(j)}]^T$, $j = 1(1)m$, should satisfy

$$x_n^{(j)} = 0(1), \qquad y_n^{(j)} = 0(\epsilon) \quad \text{as} \quad \epsilon \to 0, \qquad n = 1(1)T/\tau. \tag{3.4}$$

DEFINITION 3.1

Suppose we are given a method (2.3) which is ϵ-bounded on a class of problems of type (3.1). We then call this method ϵ-accurate on this class, if in relations (2.4) an $0(\epsilon)$ behaviour appears for all y-components of $\Phi^{(j)}$. □

Clearly, if a method is ϵ-accurate it can be used to generate approximation sequences satisfying (3.4). The next theorem summarizes the main results we obtained for method (1.2) when applied to class (3.1)[2]:

THEOREM 3.2.

(i) *Any Rosenbrock method* (1.2) *is* ϵ*-bounded on the two classes of problems* (3.1) *for which, respectively,* $A = 0$ *and* A, μ *are constant.*

(ii) *Any Rosenbrock method* (1.2) *is* ϵ*-bounded on the whole class* (3.1), *if at each stage* $J(t, x)$ *is evaluated at the special point* $(t, X) = (t^{(j)}, X^{(j)})$.

(iii) *Any Rosenbrock method* (1.2) *is* ϵ*-accurate on the whole class* (3.1), *iff the stability function* $R^{(m)}(z)$, *as well as all internal stability functions* $R^{(j)}(z)$, $j < m$, *have a zero at infinity.*

(iv) *Any Rosenbrock method* (1.2) *evaluating* $J(t, x)$ *once per step is* ϵ*-bounded on the whole class* (3.1), *iff* $R^{(j)}(\infty) = 0$ *for* $j < m$.

(v) *Consider class* (3.1). *Let the point* $X = (x, y)$ *occurring in Definition* 2.1 *be such that* $x = 0(1)$, $y = 0(\epsilon)$. *Then any Rosenbrock method* (1.2) *is* ϵ*-accurate on the whole class* (3.1). □

REMARK 3.1.

As shown in [2], ϵ-boundedness of (1.2) with respect to (3.1) is determined by the boundedness for $\epsilon \in (0, \epsilon_0]$ of

$$\epsilon^{-1} A(t) y + \gamma\tau\epsilon^{-2} A(\hat{t})[I - \gamma\tau\epsilon^{-1}\mu(\hat{t})B]^{-1}\mu(t)By, \qquad t \neq \hat{t}. \tag{3.5}$$

We shall use this rule to select an appropriate example model for the experiments. It is needed because for a specific example the conditions of Theorem 3.2 may happen to be too strong. □

3.2 *A non-linear test example*

We consider the system

$$\begin{aligned} \dot{x}_1 &= a_1(x_1 + x_2 + y - 1)^k + \epsilon^{-1}\mu_1(t)y, \\ \dot{x}_2 &= a_2(x_1 + x_2 + y - 1)^k + \epsilon^{-1}\mu_2(t)y, \\ \dot{y} &= a_3(x_1 + x_2 + y - 1)^k - \epsilon^{-1}\mu(t)y. \end{aligned} \tag{3.6}$$

Here $t_0 \le t \le T$ and $\epsilon \in (0, \epsilon_0]$, $x_1(t), x_2(t), y(t)$ are scalar, a_i and $k \ge 1$ are constant, and $\mu = \mu_1 + \mu_2$. The sum $s = x_1 + x_2 + y$ satisfies

$$\dot{s} = a(s-1)^k, \qquad a = a_1 + a_2 + a_3, \tag{3.7}$$

so that

$$s(t) = 1 + [a(1-k)t + C]^{1/(1-k)}, \qquad C = (s(t_0) - 1)^{1-k} - a(1-k)t_0. \tag{3.8}$$

If $s(0) = s_0 \neq 1$, equation (3.6) thus possesses a unique solution which is bounded on any finite interval $[0, T]$ uniformly in $\epsilon \in (0, \epsilon_0]$. Furthermore, this solution satisfies the inequalities in Theorem 3.1.

Elaborating expression (3.5) for system (3.6) yields

$$\epsilon^{-1}\mu_i(t)y\left\{\frac{\mu_i(t) + \gamma\tau\epsilon^{-1}[\mu_i(t)\mu(\hat{t}) - \mu_i(\hat{t})\mu(t)]}{\mu_i(t)[1 + \gamma\tau\epsilon^{-1}\mu(\hat{t})]}\right\}, \qquad i = 1, 2. \tag{3.9}$$

If $\mu_i(t)\mu(\hat{t}) \neq \mu_i(\hat{t})\mu(t)$, this expression is not bounded in $\epsilon \in (0, \epsilon_0]$, i.e. the conditions of Theorem 3.2 apply to the specific example (3.6). If $\mu_i(t)\mu(\hat{t}) = \mu_i(\hat{t})\mu(t)$ for all $t, \hat{t} \in [0, T]$, any Rosenbrock method (1.2) is able to generate approximation sequences bounded in $\epsilon \in (0, \epsilon_0]$.

The eigenvalues of the Jacobian $\partial F(t, X, \epsilon)/\partial X$, evaluated on the exact solution, are given by

$$\delta_1 = 0, \qquad \delta_2 = a\theta^{-1}(t), \qquad \delta_3 = -\epsilon^{-1}\mu(t), \tag{3.10}$$

where $\theta(t) = k^{-1}[at(1-k) + C]$. In the following we, therefore, take $C > 0$ and $a < 0$, so that $\delta_2 < 0$. Note that δ_2 does not depend on ϵ.

Obviously, much freedom is left in choosing the various defining parameters in (3.6). We put ($\mu = \mu_1 + \mu_2$)

$$k = 2, \qquad a_1 = -0.1, \qquad a_2 = 1, \qquad a_3 = -1, \qquad \mu_1(t) = e^t - t, \qquad \mu_2 = t. \tag{3.11}$$

Note that for all $t, \hat{t}$ we have $\mu_i(t)\mu(\hat{t}) \neq \mu_i(\hat{t})\mu(t)$, $i = 1, 2$. Further, as $\mu_2(0) = 0$, x_2 has no transient. It remains to choose a range of ϵ-values and initial values at $t = 0$. The ϵ-range will be given below with the actual experiments. Here we define two sets of initial values, namely

$$x_1(0) = 0, \qquad x_2(0) = 1, \qquad y(0) = \frac{1}{4}, \tag{3.12a}$$

$$x_1(0) = \frac{1}{4}, \qquad x_2(0) = 1, \qquad y(0) = \epsilon. \tag{3.12b}$$

The initial values (3.12b) define a smooth solution ($y(0) = \epsilon$).

3.2 *Numerical illustration*

The lack of ϵ-boundedness, or ϵ-accuracy, manifests itself by unusually large errors and, typically, the smaller ϵ, the larger the errors. We shall illustrate this undesirable phenomenon for the problems (3.6), (3.11), (3.12a) and (3.6), (3.11), (3.12b).

For the experiments we selected four simple 2-stage formulas (1.2) of order 2. All are L-stable and $R^{(1)}$ and $R^{(2)}$ are given by ($\gamma_0 = \gamma_1 = \gamma$)

$$R^{(1)}(z) = \frac{1 + (\lambda_{10} - \gamma)z}{1 - \gamma z}, \qquad R^{(2)}(z) = \frac{1 + (1 - 2\gamma)z}{(1 - \gamma z)^2}, \qquad \gamma = 1 - \frac{1}{2}\sqrt{2}. \tag{3.13}$$

Note that the formulas share the stability function $R^{(2)}$. We have $\lambda_{20} = 1 - \lambda_{21}$, $\nu_1 = 1/2\lambda_{21}$ and $\lambda_{21} = (\frac{1}{2} - \gamma)/\lambda_{10}$:

Formula	λ_{10}	α_{00}	α_{10}	α_{11}	$R^{(1)}(\infty)$	$J(t, X)$	ϵ-bounded on (3.1)	ϵ-accurate on (3.1)
a	$1-2\gamma$	1	1	0	$\frac{3\gamma-1}{\gamma}$	1	no	no
b	γ	1	1	0	0	1	yes	yes
c	$1-2\gamma$	1	0	1	$\frac{3\gamma-1}{\gamma}$	2	yes	no
d	γ	1	0	1	0	2	yes	yes

(3.14)

The choice $\lambda_{10} = \gamma$ implies $R^{(1)}(\infty) = 0$. The choice $\lambda_{10} = (3\gamma - 1)/\gamma$ is, for our purpose, rather arbitrary. Of importance is that in this case $R^{(1)}(\infty) \neq 0$. The present λ_{10}-value implies $\nu_1 = 1$ and $R^{(1)}(\infty) \simeq -0.4$. The $\alpha_{j,l}$-values are self-evident. Recall that schemes using more than one $J(x, X)$-evaluation per step are usually not recommended.

In Figs. 1–4 we plotted, for a set of ϵ-values from the interval $[10^{-7}, 1]$, the numbers $ac_x = -^{10}\log$ (max. abs. error of x-components) and $ac_y = -^{10}\log$ (abs. error of y-component) for precisely one integration step of length 1/20. We deliberately did not give errors measured after a number of steps because we noticed cancellation of x_1-errors and x_2-errors when performing more than one step. For our purpose it suffices to consider only one step. Recall that problem (3.6), (3.11), (3.12a) exhibits a transient behaviour, whereas the solution of (3.6), (3.11), (3.12b) is smooth due to the initial value $y(0) = \epsilon$.

Let us first discuss the results for (3.6), (3.11), (3.12a). Figure 1 clearly shows the lack of ϵ-boundedness of scheme a, i.e. for increasing stiffness its accuracy strongly decreases, whereas the accuracy of b and c remains constant. Also note that, in this case, scheme d is much more accurate than b and c. Figure 1 shows that d even takes advantage of increasing stiffness (this phenomenon cannot be explained from the notions of ϵ-boundedness and ϵ-accuracy). Figure 2 clearly shows the lack of ϵ-accuracy of scheme a. It should be noted that scheme c, which according to (3.14) is not ϵ-accurate, yields the same initial y-errors as b and d. This can be explained from the following (heuristic) observation. Consider the linear part of the third component of equation (3.6), i.e. $\dot{y} = -\epsilon^{-1}\mu(t)y$. Application of the 2-stage schemes c and d to this equation yields

$$\frac{y_{n+1}}{y_n} = \frac{1 - (\lambda_{20} - \gamma)\tau\epsilon^{-1}\mu(t_n) - (\lambda_{21} - \gamma)\tau\epsilon^{-1}\mu(t_n + \nu_1\tau)}{(1 - \gamma\tau\epsilon^{-1}\mu(t_n))(1 - \gamma\tau\epsilon^{-1}\mu(t_n + \nu_1\tau))} = 0(\epsilon), \qquad \epsilon \to 0. \tag{3.15}$$

Hence the extra Jacobian evaluation yields extra damping, even if $y_n^{(1)}/y_n = 0(1)$.

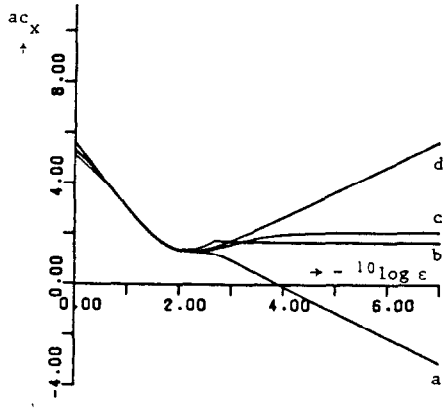

Fig. 1. Initial x-error, (3.6), (3.11), (3.12a).

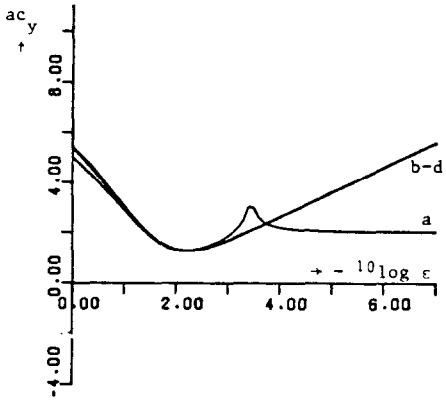

Fig. 2. Initial y-error, (3.6), (3.11), (3.12a).

The results for the easier problem (3.6), (3.11), (3.12b) have been plotted in Figs. 3 and 4. For this problem all x-approximations are 0(1) and all y-approximations are $0(\epsilon)$ (cf. Theorem 2, part (v)). Note however that the ϵ-bounded schemes b–d yield significantly more accuracy than scheme a. Finally it is worthwhile to observe that for the larger ϵ-values, say $\epsilon \in [10^{-2}, 1]$, all four schemes yield approximately the same errors.

4. MODEL CLASS 2

The second model class we are interested in, and which was also discussed in the previous paper[2], is reviewed in Section 4.1. Section 4.2 deals with a specific example which is used in Section 4.3 for a numerical illustration.

4.1 *The class of D-stability model equations*

The following class of linear stiff model problems, class $\mathcal{S}$, was proposed by Van Veldhuizen[4] (cf. (2.2)):

$$\dot{X} = F(t, \epsilon)X = \epsilon^{-1} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} X, \quad X(t) \in C^2, \tag{4.1}$$

where (a) $a_{ij} \in C$ depends smoothly on $t \in [0, T]$ and $\epsilon \in (0, \epsilon_0]$. (b) $F(t, \epsilon) = E(t, \epsilon)D(t, \epsilon)E^{-1}(t, \epsilon)$, where

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & \epsilon^{-1}d_2 \end{bmatrix}, \qquad Re(d_2(t, \epsilon)) \leq \tilde{d}_2 < 0 \text{ on } [0, T] \times (0, \epsilon_0].$$

d_1, d_2, E and E^{-1} depend smoothly on t, ϵ and the derivatives from order zero up to a sufficiently high order are bounded on $[0, T] \times (0, \epsilon_0]$.

Van Veldhuizen used class $\mathcal{S}$ in his D-stability investigations. Though presented in a somewhat different setting D-stability may be viewed as a uniform boundedness property, like ϵ-boundedness. However, it only applies to linear homogeneous problems $\dot{X} = F(t)X$. For reasons of presentation we, therefore, do not make use of van Veldhuizen's definition which is slightly different from ours (see [2, 4]).

As pointed out in [4], a nice feature of model (4.1) is the possibility of defining subclasses of $\mathcal{S}$ which describe certain types of couplings between smooth and stiff solution components. Because these couplings may be of decisive importance for the performance of a Rosenbrock method, we give a short description of these subclasses. Consider a problem from class $\mathcal{S}$. Denote $Y(t) = E^{-1}(t)X(t)$. Then Y satisfies

$$\dot{Y} = [D(t) - C(t)]Y, \qquad C(t) = E^{-1}(t)\dot{E}(t). \tag{4.2}$$

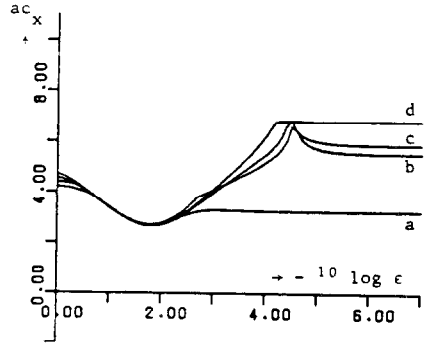

Fig. 3. Initial x-error, (3.6), (3.11), (3.12b).

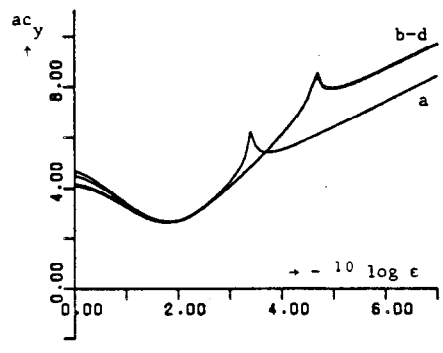

Fig. 4. Initial y-error, (3.6), (3.11), (3.12b).

In case $C(t)$ is diagonal on $[0, T]$, the problem from $\mathcal{S}$ has been uncoupled by the transformation $X = EY$, i.e. there exists no coupling between smooth and transient components. Otherwise we employ

DEFINITION 4.1

The coupling from the smooth to the transient component, at $t = t^*$, is weak if $C_{21}(t^*) = 0(\epsilon)$. The coupling from the transient to the smooth component, at $t = t^*$, is weak if $C_{12}(t^*) = 0(\epsilon)$. If a coupling is not weak, we call it strong. $\mathcal{W}_{st}(\mathcal{W}_{ts})$ denotes the subclass of $\mathcal{S}$ for which on the whole time interval $C_{21}(t) = 0(\epsilon)(C_{12}(t) = 0(\epsilon))$. □

Due to assumptions (4.1a, b) the matrix $C(t)$ is at least 0(1) as $\epsilon \to 0$. Hence problem (4.2) is of type (3.1). By means of Theorem 3.1, and the bounded transformation $X = EY$, it thus follows that all solutions of (4.1) are bounded in $\epsilon \in (0, \epsilon_0]$.

THEOREM 4.1.

Consider an arbitrary 2-stage Rosenbrock method (1.2) *which evaluates* $J(t, X)$ *once per integration step. This method is* (i) *ϵ-bounded on $\mathcal{W}_{ts}$.* (ii) *ϵ-bounded on $\mathcal{W}_{st}$, iff $R^{(1)}(\infty) = 0$.* (iii) *not ϵ-bounded on $\mathcal{S}$.*

Proof. This theorem is a special case of Theorem 3.1 in [4]. □

THEOREM 4.2.

An m-stage Rosenbrock method (1.2) *is ϵ-bounded on $\mathcal{S}$ iff at each stage $J(t, X)$ is evaluated at the special point $(t, X) = (t^{(j)}, X^{(j)})$.*

Proof. The necessity follows from Theorem 4.1, part (iii). Recall that boundedness of the m-stage result implies, by definition, boundednesss of the preceding $(m-1)$-stage results. The sufficiency has been proved in [2], Theorem 3.1. □

These two theorems show that if we have a strong coupling from stiff to smooth, and vice versa, ϵ-boundedness cannot be guaranteed if we restrict ourselves to one $J(t, x)$-evaluation per integration step. Unfortunately, schemes which reevaluate the Jacobian every stage are usually not recommended because of their considerable computational overhead.

So far we have not yet attempted to prove (i) and (ii) of Theorem 4.1 for methods (1.2) using more than 2 stages. We do conjecture, however, that these methods are also ϵ-bounded on $\mathcal{W}_{ts}$, and ϵ-bounded on $\mathcal{W}_{st}$, *iff* $R^{(j)}(\infty) = 0$, $j < m$. For example, the class consisting of all problems

$$\dot{X} = \begin{bmatrix} a_{11} & \epsilon^{-1}a_{12} \\ a_{21} & \epsilon^{-1}a_{22} \end{bmatrix} X \tag{4.3}$$

satisfying properties (4.1a, b), is a subclass, say $\mathcal{S}_2$, of $\mathcal{W}_{st}$ [2]. Because (4.3) may also be viewed as a prototype of the first variational form of model (3.1), part (iv) of Theorem (3.2) applies. It thus follows that an m-stage Rosenbrock method (1.2), using one $J(t, X)$-evaluation per step, is ϵ-bounded on $\mathcal{S}_2$, *iff* $R^{(j)}(\infty) = 0$ for $j < m$.

Because $\mathcal{S}_2 \subset \mathcal{W}_{st}$, class $\mathcal{S}_2$ does not describe strong couplings from smooth to transient. This fact may be considered as a shortcoming of equation (4.3), and thus also of (3.1), when used as a model.

4.2 *A test example exhibiting only strong couplings*

Consider the problem (see also [2, 7, 8])

$$\dot{X} = E(t)\begin{bmatrix} d_1(t) & 0 \\ 0 & \epsilon^{-1}d_2(t) \end{bmatrix} E^{-1}(t)X, \qquad E(t) = \begin{bmatrix} \cos\theta t & -\sin\theta t \\ \sin\theta t & \cos\theta t \end{bmatrix}, \tag{4.4}$$

θ being constant. Then $Y(t) = E^{-1}(t)X(t)$ satisfies (cf. (4.2))

$$\dot{Y} = \begin{bmatrix} d_1(t) & \theta \\ -\theta & \epsilon^{-1}d_2(t) \end{bmatrix} Y. \tag{4.5}$$

Hence $C_{12}(t) = -\theta$, $C_{21}(t) = \theta$. Consequently, we have to deal with a strong coupling from stiff to

smooth, and vice versa. It is not difficult to verify that for this specific example part (iii) of Theorem 4.1 applies. Note that equation (4.5) belongs to $\mathcal{S}_2$. Let $d_1 = d_2 = -1$. Then

$$Y(t) = \begin{bmatrix} 1+\epsilon\lambda^+ & 1+\epsilon\lambda^- \\ -\epsilon\theta & -\epsilon\theta \end{bmatrix} \begin{bmatrix} C^+ e^{\lambda^+ t} \\ C^- e^{\lambda^- t} \end{bmatrix}, \tag{4.6}$$

where $C^{\pm}$ are arbitrary constants and $\lambda^{\pm} = (1/2)(-1-\epsilon^{-1} \pm \sqrt{[(1-\epsilon^{-1})^2 - 4\theta^2]})$. Note that $\lambda^- \sim -\epsilon^{-1}$ and $\lambda^+ \to -1$ as $\epsilon \to 0$. Next we set $C^- = 0$, $C^+ = 1$. Then

$$X(t) = E(t) \begin{bmatrix} (1+\epsilon\lambda^+) e^{\lambda^+ t} \\ -\epsilon\theta\, e^{\lambda^+ t} \end{bmatrix} = \begin{bmatrix} e^{-t}\cos\theta t \\ e^{-t}\sin\theta t \end{bmatrix} + 0(\epsilon), \qquad \epsilon \to 0. \tag{4.7}$$

We see that the solution (4.7) is smooth and, to a great extent, independent of the stiffness parameter ϵ. The same remark applies to the first component of the corresponding solution of (4.5). Its second component is $0(\epsilon)$. In what follows we shall refer to the X-example and Y-example.

4.3 *Numerical illustration*

We integrated the X-example and Y-example for $\theta = 1$ and for a set of ϵ-values from $[10^{-8}, 10^{-1}]$ with all four 2-stage formulas (3.14) over the t-interval $[0, 2\pi]$, using a constant stepsize $\tau = \pi/25$. The initial values at $t = 0$ are defined in equations (4.6), (4.7). Note that for the Y-example the formulas (3.14) are identical.

In Fig. 5 we plotted the value $ac = -^{10}\log$ (max. abs. error at $t = 2\pi$) against ϵ. The a-curve and b-curve clearly show the lack of ϵ-boundedness of methods a and b (instability for small ϵ). Methods c and d are ϵ-bounded on $\mathcal{S}$ (see Theorem 4.2). They produce approximations which are nearly independent of ϵ. Recall that, for small ϵ, the exact solutions share this property. Finally, this example nicely shows that a simple transformation of the differential equation may lead to a qualitatively different behaviour of a Rosenbrock method.

5. THE AUTONOMOUS FORM

Many authors prefer the autonomous form. It facilitates the analysis of the consistency conditions, while every non-autonomous equation (1.1) can be converted to the autonomous form by introducing t as a new dependent variable. For example, the Rosenbrock code ROW4A requires the autonomous form [5]. When we rewrite problem (1.1) in the autonomous form the derivative F_t enters into the computation. It is easily seen that the Rosenbrock approximation (1.2) then can be defined by the (non-autonomous) scheme

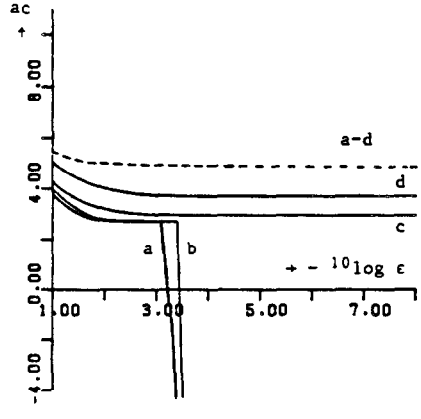

Fig. 5. —— X-example, ---- Y-example.

$$
\begin{aligned}
&X_n^{(0)} = X_n,\\
&K_n^{(j)} = [I - \gamma_j \tau J_n^{(j)}]^{-1}[F(t_n^{(j)}, X_n^{(j)}) + \gamma_j \tau G_n^{(j)}], \qquad j = 0(1)m-1,\\
&X_n^{(j)} = X_n + \tau \sum_{l=0}^{j-1} \lambda_{j,l} K_n^{(l)}, \qquad j = 1(1)m,\\
&X_{n+1} = X_n^{(m)},
\end{aligned}
\tag{5.1}
$$

where $G_n^{(j)} = G(\hat{t}_n^{(j)}, \hat{X}_n^{(j)})$, $G(t, X) = \partial F(t, X)/\partial t$. Furthermore, $t_n^{(j)}$ is now defined by $t_n^{(j)} = t_n + \tau(\lambda_{j,0} + \cdots + \lambda_{j,j-1})$. All other quantities are defined as in scheme (1.2). It is convenient to use the notation (5.1) (cf. [9, 10]).

Because we deal with non-autonomous models, the following interesting question arises. When we apply (5.1) to the model classes (3.1) and (4.1), do we then preserve the boundedness results summarized in the two preceding sections? For the most interesting results the answer to this question is, peculiarly, negative. It is even negative for schemes using more than one Jacobian evaluation per step. This matter will be discussed in Section 5.1. By way of illustration, we also repeat the experiments presented before.

5.1 *Boundedness results for method* 5.1

THEOREM 5.1

No Rosenbrock method (5.1) *is* ϵ*-accurate on class* (3.1).

Proof. By counterexample. Consider the simplified problem $\dot{y} = -\epsilon^{-1}\mu(t)y$. Application of any method (5.1) at a point (t, y) yields

$$
y^{(1)} = \frac{1 - (\lambda_{10} - \gamma_0)\tau\epsilon^{-1}\mu(t) - \lambda_{10}\gamma_0\tau^2\epsilon^{-1}\dot{\mu}(t)}{1 + \gamma_0\tau\epsilon^{-1}\mu(t)}\, y. \tag{5.2}
$$

We see that if $(\lambda_{10} - \gamma_0)\mu(t) + \lambda_{10}\gamma_0\tau\dot{\mu}(t) \neq 0$, then $y^{(1)} = 0(1)$ as $\epsilon \to 0$. By definition, ϵ-accuracy of an m-stage method implies $y^{(1)} = 0(\epsilon)$. □

If we apply relation (5.2) repeatedly, we may easily encounter instability. For example, substituting $\lambda_{10} = \gamma_0$ (L-stability) and $\mu(t) = \exp(-t/\tau\gamma_0)$ yields $y^{(1)} = y$. On the other hand, when using the non-autonomous form the substitution $\lambda_{10} = \gamma_0$ yields $|y^{(1)}/y| = |(1 + \gamma_0\tau\epsilon^{-1}\mu(t))^{-1}| < 1$ for all $\tau > 0$ and $\epsilon \in (0, \epsilon 0]$. In other words, the stability of the 1-stage scheme may be lost by conversion to the autonomous form. Without doubt this conclusion also applies to m-stage schemes, $m > 1$. As we do not discuss stability properties we do not pursue this subject further.

THEOREM 5.2

(i) *No method* (5.1) *is* ϵ*-bounded on class* $\mathcal{S}_2$. *Consequently, no method* (5.1) *is* ϵ*-bounded on class* $\mathcal{S}$ *and class* (3.1).

(ii) *Any method* (5.1) *is* ϵ*-bounded on the two classes of problems* (3.1) *for which, respectively,* $A = 0$ *and* A, μ *are constant.*

(iii) *Consider class* (3.1). *Let the point* $X = (x, y)$ *occurring in Definition* (2.1) *be such that* $x = 0(1)$ *and* $y = 0(\epsilon)$. *Then any Rosenbrock method* (5.1) *is* ϵ*-accurate on the whole class* (3.1).

Proof. The proofs of (ii)–(iii) go along the same lines as the proofs of the corresponding parts of Theorem 3.2 (see [2], Section 4). The proof of part (i) goes by counterexample. It suffices to take $m = 1$. Consider the problem (cf. (4.3))

$$
\begin{aligned}
\dot{x} &= \epsilon^{-1} a_{12}(t) y,\\
\dot{y} &= -\epsilon^{-1} y.
\end{aligned}
\tag{5.3}
$$

The 1-stage scheme, applied at a point (t, x, y), yields the increment vector

$$K^{(0)} = \begin{bmatrix} \epsilon^{-1} a_{12}(t)(1 + \gamma_0 \tau \epsilon^{-1})^{-1} y + \gamma_0 \tau \epsilon^{-1} \dot{a}_{12}(t) y \\ -\epsilon^{-1}(1 + \gamma_0 \tau \epsilon^{-1}) y \end{bmatrix}. \tag{5.4}$$

By an appropriate choice of $a_{12}(t)$, the first component becomes unbounded in $\epsilon \in (0, \epsilon_0]$. This simple observation proves part (i). □

By way of illustration we repeated the aforementioned experiments with the four 2-stage schemes (3.14), but now using the autonomous form. Figures 6–10 correspond to Figs. 1–5, respectively. Note that all four schemes now behave more or less the same.

5.2 *An experiment with ROW4A*

ROW4A is an automatic Rosenbrock code based on the algorithm GRK4A published in [11]. Gottwald and Wanner [5] provided it with a so-called backstep strategy to obtain a more reliable stepsize and local error control. The underlying integration method is A-stable and of order 4. Its increment vectors are of the (more general) form (cf. (1.2))

$$K_n^{(j)} = [I - \gamma \tau J_n]^{-1} \left\{ F(X_n^{(j)}) + \sum_{l=0}^{j-1} \beta_{j,l} K_n^{(l)} \right\}, \tag{5.5}$$

and assume the autonomous form ((5.5) can also be rewritten like formula (5.1), see [9]). The method is not ϵ-bounded on class (3.1) and class $\mathcal{S}$. It uses three $F(X)$-evaluations and one

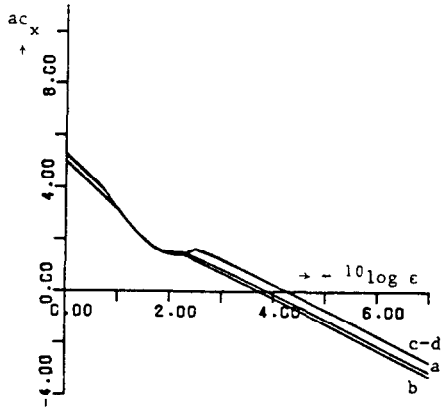

Fig. 6. Initial x-error, (3.6), (3.11), (3.12a), Autonomous form.

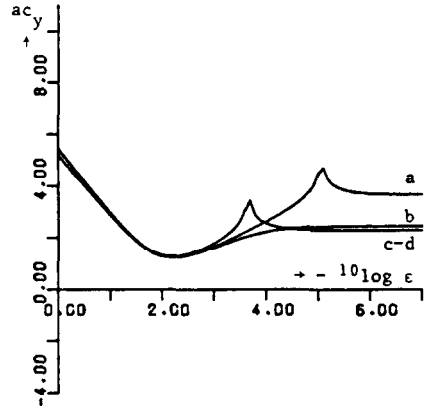

Fig. 7. Initial y-error, (3.6), (3.11), 3.12a), Autonomous form.

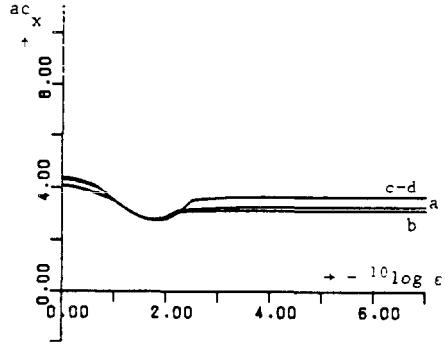

Fig. 8. Initial x-error, (3.6), (3.11), (3.12b), Autonomous form.

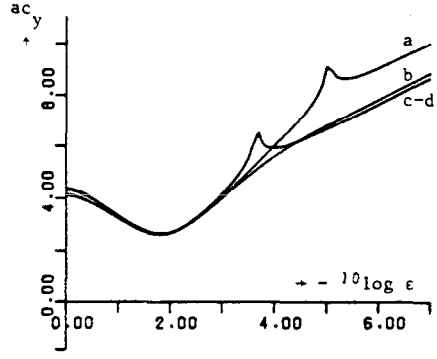

Fig. 9. Initial y-error, (3.6), (3.11), (3.12b), Autonomous from.

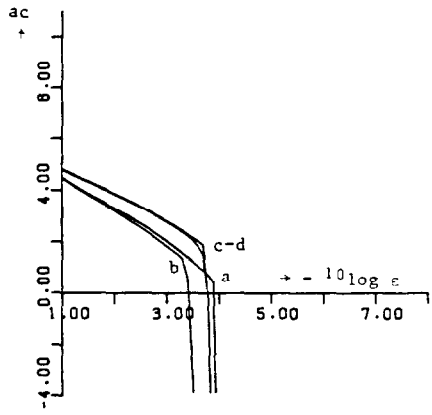

Fig. 10. X-example; Autonomous form.

$J(X)$-evaluation per step. We implemented ROW4A on a CDC Cyber 750 in single precision (14 decimals). Our version computes J_n from the analytic expression.

Our aim of reporting an experiment with an automatic code, like ROW4A, is to illustrate how the lack of ϵ-boundedness shows up in practice. When this property is missing, one may encounter unusually large local errors, even when the solution to be integrated is smooth. The local error control mechanism of a reliable automatic code will detect these too large errors and will, at the cost of more integration steps of course, reduce the stepsize to an appropriate level. Thus, roughly speaking, in practice the lack of ϵ-boundedness shows up in the number of integration steps (see also [4], Section 5).

The experiment consists of the automatic integration of the X-example and Y-example of Section 4.3, over the interval $[0, 2\pi]$, for a set of ϵ-values between 10^{-7} and 10^{-1}. The tolerance parameter TOL of ROW4A and the initial stepsize were in all integrations equal to 10^{-3} and 10^{-2}, respectively. TOL is used in a local error test at each integration step. The local error is estimated by means of an embedded formula of order 3. More specifically, the maximum norm of the difference vector of the 4th order and 3rd order result, multiplied by a certain scaling factor, is used for measuring the estimation of the local error. At each step this quantity is kept smaller than TOL (see [5, 11] for details). If necessary, the code rejects the integration step and continues with a reduced stepsize. Figure 11 shows results of the experiment.

The plots clearly show the lack of ϵ-boundedness of ROW4A when applied to the X-example. Though the exact solution is smooth, and nearly independent of ϵ, the numbers IPAS and IREP strongly increase as ϵ decreases. As observed above, such a behaviour was to be expected. However, a more dramatic observation is that the code loses its accuracy. The global error strongly increases as ϵ decreases. For the smaller ϵ-values the solution delivered by the code is completely wrong. This experiment demonstrates that it can be very dangerous to rely on local error control mechanisms.

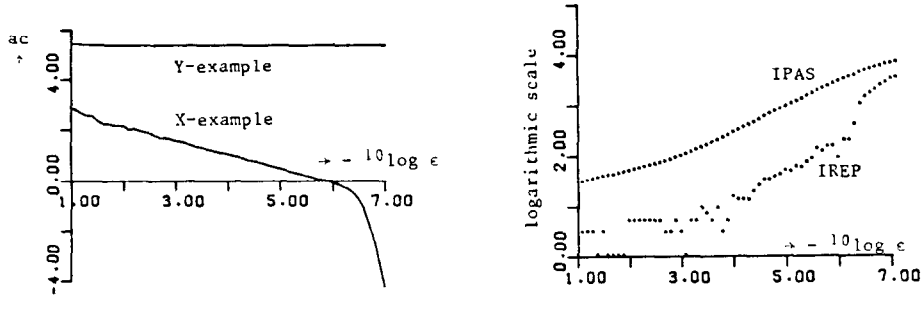

Fig. 11. Results for ROW4A. In the right figure we plotted IPAS = the number of accepted steps and IREP = the number of repeated steps needed by ROW4A on the X-example. For the Y-example these numbers are 16 and 0, respectively, and do not change with ϵ.

6. SOME FINAL COMMENTS

The question arises as to how to employ our experience in order to improve the Rosenbrock methods when applied to real life problems. Let us first consider methods based on the *non-autonomous* form (11). For this type of Rosenbrock method our results strongly suggest taking care of ϵ-boundedness and ϵ-accuracy when dealing with problems where the stiffness originates from t-dependent parts in the equation. However, if one wishes to construct such a method, one has to face an additional difficulty, i.e. the solution of extra order conditions due to the presence of derivatives with respect to t. To solve these extra conditions, for a given order, it may well be necessary to add extra stages. From this point of view the *autonomous* form should be preferred. Unfortunately, for the type of problems mentioned above, *the conversion to the autonomous form* may lead to a significant loss in accuracy, as shown in our experiments. This circumstance makes it difficult to decide which approach should be preferred. In the author's opinion, a justifiable decision can be made only if one has a typical problem class at hand. In this connection we should also remark that Kaps and Rentrop[11] and Gottwald and Wanner[5] report promising results with their "autonomous" codes GRK4A and ROW4A. Gottwald and Wanner[12] even show that on a set of four real life problems from chemical kinetics and physiology, their code ROW4A is more efficient and more reliable than a popular backward differentiation one.

Acknowledgement-The author would like to express his gratutide to Mrs. M. Louter-Nool for her assistance in preparing the plots.

REFERENCES

1. H. H. Rosenbrock, Some general implicit processes for the numerical solution of differential equations. *Computer J.* **5**, 329–330 (1963).
2. J. G. Verwer, An analysis of Rosenbrock methods for non-linear stiff initial value problems. *SIAM J. Numer. Anal.*, to appear, 1982.
3. H. J. Stetter, Towards a theory for discretizations of stiff differential system. *Lecture Notes in Mathematics* 506, pp. 190–201. Springer Verlag, Berlin (1976).
4. M. Van Veldhuizen, D-stability. *SIAM J. Numer. Anal.* **18**, 45–64 (1981).
5. B. A. Gottwald and G. Wanner, A reliable implementation of one-step methods for differential equations. *Computing* **26**, 355–360 (1981).
6. E. Griepentrog, Numerische Integration stiefer Differentialgleichungs-systeme mit Einschrittverfahren. *Beiträge zur Numerischen Mathematik* **8**, 59–74 (1980).
7. H. O. Kreiss, Difference methods for stiff ordinary differential equations. *SIAM J. Numer. Anal.* **15**, 21–58 (1978).
8. J. Sand, *A Note on a Differential System Constructed by H. O. Kreiss.* Report TRITA-NA-8004, The Royal Institute of Technology, Stockholm, Sweden (1980).
9. P. Kaps and G. Wanner, A study of Rosenbrock type methods of high order. *Numer. Math.* **38**, 297–298 (1981).
10. L. F. Shampine, *Implementation of Rosenbrock Methods.* Report SAND80-2367J, Sandia National Laboratories, Albuquerque, New Mexico (1980).
11. P. Kaps and P. Rentrop, Generalized Runge–Kutta methods of order 4 with stepsize control for stiff ordinary differential equations. *Numer. Math.* **33**, 55–68 (1979).
12. B. A. Gottwald and G. Wanner, Stiff systems of ordinary differential equations in biology and chemistry: Validation of numerical methods for their solution. In *Continuous Simulation of Physical Systems* (Edited by T. D. Bui), to appear.